RAPORTARE ŞTIINŢIFICĂ
privind proiectul
Alcătuirea și explorarea unui corpus oral de română contemporană vorbită din perspectivă lingvistică
(Codul proiectului: PN-III-P1-1.1-PD-2019-1029)
Etapa 2 (ianuarie – decembrie 2021)

I. REZUMATUL ETAPEI

Obiectivele proiectului de față privesc configurarea si explorarea din perspectivă lingvistică a unui corpus oral de română contemporană în vederea dezvoltării unor metode noi de analiză acustică îndreptate către modernizarea cercetărilor fonetice din România. Cea de a doua etapă a proiectului, desfășurată în intervalul ianuarie – decembrie 2021, a presupus *dezvoltarea corpusului și studierea variației lingvistice*. În acest sens, au fost parcurse, conform contractului, următoarele activități:

- *înregistrări audio de calitate înaltă, înregistrare discurs spontan, înregistrare discurs controlat, construirea corpusului* – am realizat toate înregistrările în laboratorului de fonetică din cadrul Institutului de Lingvistică al Academiei Române „Iorgu Iordan – Al. Rosetti". Respectând proiectul științific, toți cei 12 vorbitori anchetați sunt reprezentativi ai dialectului muntenesc, pe baza căruia limba română standard este formată. Au fost respectate toate aspectele privind protejarea datelor personale. În ceea ce privește discursul controlat, pentru analiza vocalelor din limba română și măsurătorile privind VOT-ul consoanelor ocluzive, numărul repetițiilor pe enunț a fost mărit de la 3 (conform propunerii), la 5 repetiții. Motivația acestei alegeri constă în obținerea unui număr mai mare de date, contribuind astfel la realizarea unui studiu acustic mult mai aprofundat și mai echilibrat din punct de vedere statistic. În ceea ce privește discursul spontan, interviurile au fost organizate în jurul unor teme comune de discuție (vezi *figura 1*), facilitând astfel compararea datelor lingvistice.

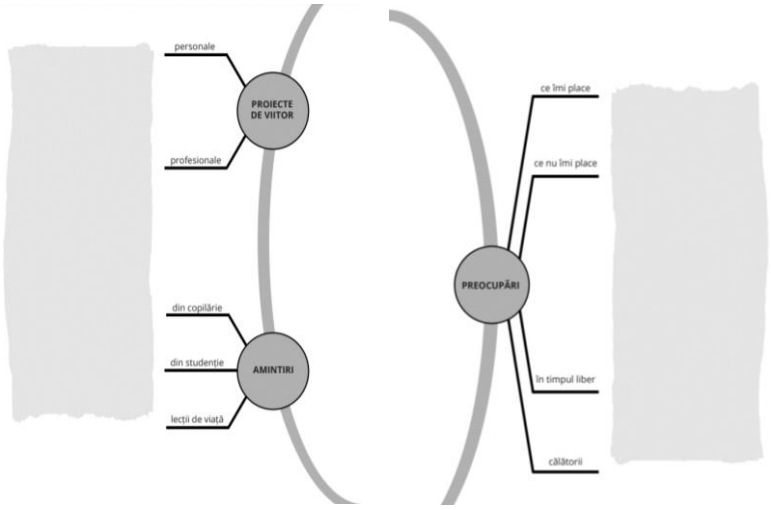


Figura 1. Model interviu discurs spontan

- *transcrierea și procesarea corpusului* – corpusul înregistrat a fost transcris și procesat. Am transcris ortografic și fonologic fișierele audio. În plus față de propunerea științifică, am optat pentru introducerea unor marcaje în transcriere (ezitări, trunchieri, elidări, repetiții, schimbare de cod, printre altele), care să contribuie la o recuperare mult mai rapidă a datelor necesare și să conducă, astfel, la o experiență de lucru mult mai bună din perspectiva utilizatorilor. Transcrierea fonologică a fost adaptată pentru a permite distincția între semivocale și semiconsoane, facilitând, pe de o parte, adnotarea corpusului și stabilirea corespondentelor sunet – grafem, iar, pe de altă parte, contribuind la o descriere fonotactică mult mai exactă a discursului oral.
- *analiza în Praat, Excel și R, vizualizarea datelor* – am utilizat diferite soft-uri pentru analiza cantitativă și calitativă a materialului înregistrat și transcris. Programul *Praat* a fost utilizat în diferite etape ale dezvoltării corpusului, mai exact de la înregistrări, la transcrieri și segmentări. De asemenea, analizele de ordin acustic, atât pentru discursul controlat, cât și cel spontan, au fost realizate în cadrul aceluiași program. Principalele măsurători acustice au fost axate pe durată, frecvența fundamentală (f0) și primii doi formanți (F1 și F2), ca de exemplu:

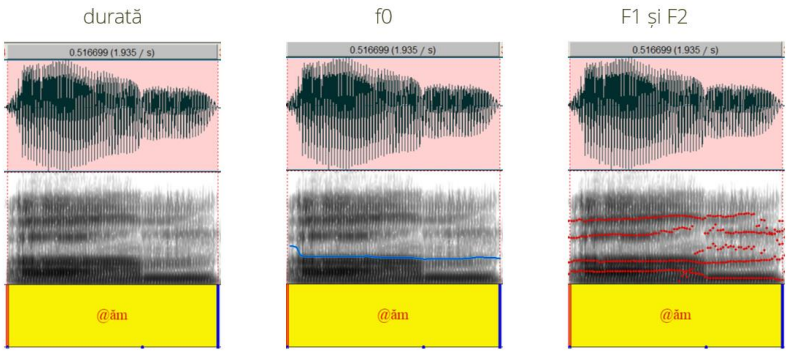
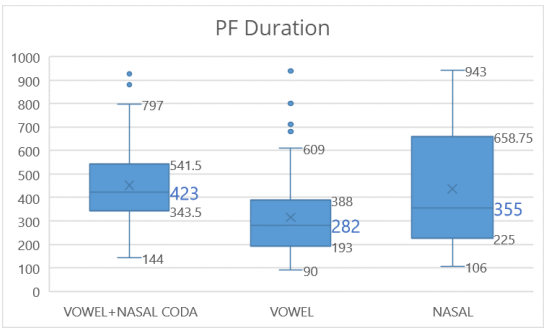


Figura 2. Analiza în Praat

Pentru analiza statistică exploratorie a datelor am optat pentru *Excel* și extensia acestuia *Kutools for Excel*. În ceea ce privește vizualizarea rezultatelor, am utilizat limbajul de programare *R* și interfața *RStudio*. De exemplu, pentru generarea spațiilor vocalice (obținute prin extragerea primilor doi formanți, vezi etapa precedentă) am folosit librăria *ggplot2* și am conceput codul în raport cu materialul analizat.

```
library(ggplot2)
setwd("D:/PAUSE_FILLERS")
data <-  read.table("type2.txt", sep="\t", header=TRUE)
culori <- c( "magenta", "orange")
ggplot (data, aes (x = F2, y = F1, col = TYPE, shape = TYPE))+ geom_point(size = 5, shape = 8)+ scale_color_manual (values = culori) + scale_x_reverse()+ scale_y_reverse()+ labs(x = "F2 (Hz)", y = "F1 (Hz)")+ ggtitle("Vocalic Space \n Pause Fillers")+ theme(panel.background = element_rect(fill = "white", colour = "grey50"))+ theme(panel.border = element_rect(linetype = 1, color = "black", fill = NA))+ theme(axis.text.x  = element_text(vjust = 0.1, size=15, color = "black"), axis.title.x = element_text(face="bold", size = 17))+ theme(axis.text.y  = element_text( vjust = 0.1, size=15, color = "black"), axis.title.y = element_text(face="bold", size = 16))+ theme(plot.title = element_text(size = 19, hjust = 0.5, face = "bold")) + theme(legend.text = element_text(colour="black", size = 17)) + theme(legend.title = element_text(colour="black", size = 17))+ theme(legend.key = element_rect(fill = "white", colour = "white"))
```

DURATION (ms)	
Mean	396.359375
Standard Error	10.94128222
Median	374.5
Mode	326
Standard Deviation	175.0605154
Sample Variance	30646.18407
Kurtosis	0.600674071
Skewness	0.747800409
Range	853
Minimum	90
Maximum	943
Sum	101468
Count	256
Largest(1)	943
Smallest(1)	90
Confidence Level(95.0%)	21.54678271



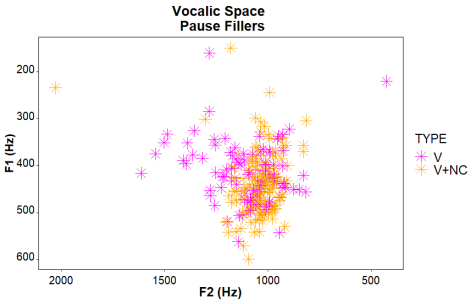


Figura 3. Analiza în Excel și R

- *descriere fonotactică, descriere de frecvență a cuvintelor în corpusul înregistrat* – am explorat corpusul transcris din perspectiva analizei cantitative a datelor. Analiza a avut la bază atât transcrierea ortografică a materialului înregistrat (unde au fost marcate, pe lângă alte fenomene, posibilele elidări din cadrul vorbirii subiecților intervievați – marcare ce nu a fost identificată și în cazul altor corpusuri disponibile online, fiind astfel o contribuție importantă a studiului nostru la descrierea variației lingvistice din limba română contemporană), cât și transcrierea fonologică adaptată tocmai pentru aceste analize cantitative (marcarea separată a semivocalelor și a semiconsoanelor; diferențierea între vocalele utilizate în contexte de ezitare sau de „particle fillers" (umplere a pauzei din discurs) și cele din cadrul lexemelor). Ne-am orientat, de asemenea, către studiul „disfluențelor" din corpusul spontan, cum ar fi repetițiile (marcate în corpus prin +), ezitările (marcate prin %), „pause fillers" (marcate prin @). Un studiu aprofundat a fost realizat pe „particle fillers", analizând atât frecvența acestor cuvinte, cât și trăsăturile lor acustice.
- *livrarea intermediară a corpusului* – am pus la dispoziția utilizatorilor fragmente din corpusului înregistrat. În acest sens, fișierele audio sunt corelate cu transcrierile ortografice și fonologice grupate în fișiere de tip TextGrid. Întrucât parcurgerea materialului disponibil pe site-ul proiectului necesită o minimă instruire tehnică din partea utilizatorului, am venit în întâmpinarea acestei necesități, în plus față de obligațiile contractuale, prin includerea unui material informativ. Sunt explicate, pas cu pas, toate etapele necesare descărcării, vizualizării și explorării corpusului. Tot în plus față de obligațiile contractuale, dar respectând principiile de bază ale proiectului nostru, anume relansarea cercetărilor plasate la interfața fonetică-fonologie, am inclus în această lansare intermediară a corpusului înregistrări cu consoanele limbii române plasate în poziție intervocalică: (1) ocluzive – /apa/ /aba/, /ata/, /ada/, /ak'a/, /ag'a/, /aka/, /aga/; (2) fricative – /afa/, /ava/, /asa/, /aza/, /aʃa/, /aʒa/, /aha/; (3) africate – /atsa/, /atʃa/, /adʒa/; (4) sonante – /ama/, /ana/, /ala/, /ara/. Toate aceste materiale dezvoltate pentru descrierea sunetelor din limba au avut numeroase aplicații (de exemplu, în cursurile de practică inițiate anul acesta de către directorul proiectului în cadrul laboratorului de fonetică; citirea spectrogramelor – vezi *figura 4*) și sunt disponibile pe site-ul proiectului.

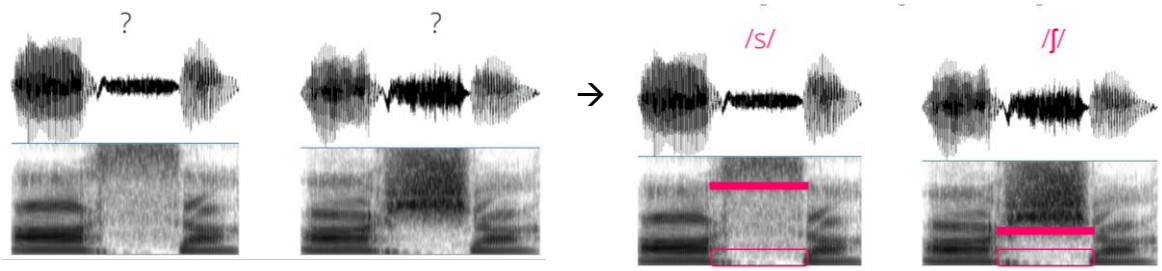


Figura 4. Citirea și interpretarea spectrogramelor (identificarea fricativelor surde)

- *efectuarea stagiului de mobilitate* – stagiu de cercetare efectuat în cadrul proiectului postdoctoral ROC-lingv (nr. 173/2020), în perioada 16 oct. – 15 nov. 2021, la Universitatea din Calabria, Catedra de Limbă și Literatură Română, Departamentul de Științe Umaniste, Italia, coordonator Prof. Danilo De Salazar. Adeverința corespunzătoare efectuării acestui stagiu de către directorul proiectului este disponibilă pe site-ul ROC-lingv.
- *documentare în cadrul institutului gazdă și al stagiului de mobilitatea* – am parcurs bibliografia de specialitate aferentă dezvoltării și explorării corpusurilor lingvistice orale, iar în cadrul stagiului de mobilitate am putut avea acces la platforme internaționale dedicate corpusurilor orale.
- *prezentarea rezultatelor proiectului la cel puțin o conferință națională sau internațională* – am prezentat 4 comunicări științifice realizate online (vezi *infra*) privind rezultatele proiectului ROC-lingv, depășind astfel numărul prevăzut inițial și asumat conform contractului. Am susținut, de asemenea, 2 prelegeri ca vorbitor invitat al Departamentului de Științe Umaniste, la Universitatea din Calabria, Italia (vezi *infra*).
- *redactarea și trimiterea spre publicare a cel puțin un articol într-o revistă de profil, națională sau internațională* – am redactat și am trimis către publicare un articol (vezi *infra*), „Developing linguistic resources for Romanian written and spoken language".
- *site-ul proiectului updatat* – am actualizat site-ul proiectului în raport cu activitățile prevăzute pentru actuala etapă de raportare.

  În ceea ce privește programul de management al proiectului, am avut ședințe cu mentorul proiectului, prof. univ. dr. Andrei Avram, am schițat obiectivele principale și intermediare ale acestei etape de raportare, asigurând astfel un flux de lucru optim.

II. DESCRIEREA ȘTIINȚIFICĂ ȘI TEHNICĂ

Scopul cercetării noastre postdoctorale constă în dezvoltarea și explorarea unui corpus oral de română contemporană vorbită. Prin crearea și punerea la dispoziție a acestui corpus publicului larg, proiectul ROC-lingv contribuie la dezvoltarea resurselor lingvistice dedicate studiului limbii române.

Am susținut 4 comunicări științifice, ca autor unic, la manifestări științifice internaționale:

(1) **Niculescu**, O. 2021. „Constructing an open-source speech corpus of contemporary standard Romanian: outline and preliminary remarks", comunicare orală realizată online în cadrul conferinței *Phonetics and Phonology: real-world applications (PaPE 2021),* la Workshop-ul *From Speech Technology to Big Data Phonetics and Phonology: a win-win paradigm*, 21 – 23 iunie 2021, Barcelona, Spania.

(2) **Niculescu**, O. 2021. „Recent tools for teaching and conducting research at the interface between phonetics and phonology in contemporary standard Romanian", comunicare orală realizată online în cadrul celui de *Al 21-lea Colocviu Internațional al Departamentului de*

Lingvistică: Orientări actuale în lingvistica teoretică și aplicată, 19 – 20 noiembrie 2021, București, România.

(3) **Niculescu**, O. 2021. „A preliminary acoustic study on pause fillers in Romanian monologue speech. Evidence from a recent developed speech corpus", comunicare orală realizată online în cadrul Conferinței internaționale anuale a Facultății de Limbi și Literaturi Străine (FLLS 2021), 26 – 27 noiembrie 2021, București, România.

(4) **Niculescu**, O. 2021. „Developing linguistic resources for Romanian written and spoken language", comunicare orală realizată online în cadrul conferinței *ConsILR 2021* (The 16th edition of The International Conference on Linguistic Resources and Tools for Natural Language Processing), 13 – 14 decembrie 2021, Iași, România.

Am susținut 2 prelegeri, ca vorbitor invitat, la Universitatea din Calabria, Departamentul de Studii Umaniste:

(5) **Niculescu**, O. 2021. „Introduction to PRAAT: Tips and Tricks for Doing Phonetics by Computer (Part 1)", vorbitor invitat, comunicare ținută la Catedra de Limbă și Literatură Română, Departamentul de Științe Umaniste, Universitatea din Calabria, 2 noiembrie 2021, Calabria, Italia.

(6) **Niculescu**, O. 2021. „Introduction to PRAAT: Tips and Tricks for Doing Phonetics by Computer (Part 2)", vorbitor invitat, comunicare ținută la Catedra de Limbă și Literatură Română, Departamentul de Științe Umaniste, Universitatea din Calabria, 4 noiembrie 2021, Calabria, Italia.

Finanțarea proiectului este recunoscută la nivelul fiecărei prezentări. Toate comunicările sunt realizate într-o limbă de circulație internațională.

Am redactat și am trimis către evaluare un articol:

Niculescu, O. 2021. „Developing linguistic resources for Romanian written and spoken language", în Proceedings of the 16th International Conference "Linguistic Resources and Tools for Processing the Romanian Language", ConsILR 2021, Iași, Romania [11 pagini predate].

În concluzie, proiectul „Alcătuirea și explorarea unui corpus oral de română contemporană vorbită din perspectivă lingvistică" (ROC-lingv), coordonat de CS dr. Oana NICULESCU în cadrul Institutului de Lingvistică al Academiei Române „Iorgu Iordan – Alexandru Rosetti", avându-l ca mentor pe prof. univ. dr. Andrei AVRAM, a îndeplinit toate obiectivele, activitățile și sarcinile prevăzute pentru cea de a doua etapă de lucru, adesea depășind prevederile contractuale.

Director Proiect,
NICULESCU OANA